DOCUMENT RESUME

ED 269 431 TM 860 261

AUTHOR Barbadora, Bernard M.
TITLE Findings of Re-Testing Six Inner-City Schools: Why and How the Cincinnati Public Schools Bite the Bullet to Assure Accuracy in Testing.
PUB DATE Apr 86
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Educational Improvement; Elementary Education; *Norm Referenced Tests; *School Districts; Scores; Statistics; *Testing; *Testing Problems; *Testing Programs; Urban Education
IDENTIFIERS Chi Square; *Cincinnati Public Schools OH; Design for School Excellence; *Retesting; T Test

ABSTRACT

This is a descriptive study of the activities, statistics, and decisions made regarding the re-testing of six inner-city elementary schools, based on results of a norm-referenced achievement test given in the Cincinnati Public School district in April, 1985. Three statistical operations were performed on test data from the Design for School Excellence, a school improvement plan, for each of the six schools. They were chi square, a "t" test using Normal Curve Equivalents (NCE) based on the school's average and a "t" test using NCE by grade in each school. Findings reveal that the obtained scores from the second testing either improved or were statistically the same as the first testing in 60 percent of the re-test situations. Therefore, results of the first testing were used and reported for the particular school/grade result in 1985. Schools where the re-test did not support the results of the first test had the initial scores invalidated and substituted for the re-test scores. In the remaining test situations substantial achievement differences were observed. Based on absolute differences, 10 of the 46 testing situations (20 percent) actually improved. Additionally both the positive and negative aspects of re-testing schools are presented. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269431

FINDINGS OF RE-TESTING SIX INNER-CITY SCHOOLS: WHY AND HOW THE CINCINNATI PUBLIC SCHOOLS BITE THE BULLET TO ASSURE ACCURACY IN TESTING

by

Bernard M. Barbadora
Testing Section, Evaluation Branch
Cincinnati Public Schools

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Barbadora

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the
American Educational Research Association,
San Francisco, California
April, 1986

TM 860 261

ABSTRACT

This is a descriptive study of the activities, statistics, and decisions made regarding the re-testing of six inner-city elementary schools, based on results of a norm-referenced achievement test given in the Cincinnati Public School district in April, 1985.

Results/findings are presented along with an extended discussion section highlighting both the positive and negative aspects of re-testing schools.

FINDINGS OF RE-TESTING SIX INNER-CITY SCHOOLS:

WHY AND HOW THE CINCINNATI PUBLIC SCHOOLS

BITE THE BULLET TO ASSURE ACCURACY IN TESTING

> "We must make it very clear to everyone including principals, teachers and community that we will leave no stone unturned to protect the integrity of the testing processes." (James N. Jacobs, 1985, late Superintendent of Cincinnati Public Schools)

The purpose of this paper is to describe in detail why the Cincinnati Public Schools do re-testing in their norm-referenced city-wide testing program, how the Cincinnati Public Schools go about re-testing, results of their most extensive re-testing effort, and a general discussion of the politics of re-testing.

There are three major reasons why the Cincinnati Public Schools are no longer only interested in, but more recently outright demanding accurate and valid test results. First, our late superintendent, James N. Jacobs, went on record with this statement, "We must make it very clear to everyone including principals, teachers and community, that we will leave no stone unturned to protect the integrity of the testing processes."

Second, a school improvement/accountability program went into operation this past school year (1984-85). By way of a little background, over the past three years the Cincinnati Public Schools have been piloting a school improvement plan now referred to as the Design for School Excellence. The plan involves the identification of merit schools, and test results are being used as the most critical output measures in identifying such schools for meritorious performance or, if you will, school excellence.

Though the third motivating factor, to assure accuracy in test results, may be less critical than the first two, it exists because of all the national attention given to the subject. Accuracy in testing, especially at AERA and NCME in 1985, received much attention and an assortment of titles such as:

CHEATING ON STANDARDIZED TESTS: WHY DOES IT EXIST? or THE EFFECTS OF FARMING ROCKY GROUND UNDER CAPRICIOUS WEATHER CONDITIONS

CHEATING - WHAT CAN WE DO ABOUT IT?

IRREGULARITIES IN TESTING: ETHICAL, PSYCHOMETRIC, AND POLITICAL ISSUES'

CHEATING ON STANDARDIZED TESTS: WHAT IS IT?

DETECTION OF COACHING AND ANSWER COPYING ON STANDARDIZED TESTS

REPORTING TEST RESULTS TO THE PUBLIC: EXPLORING THE DONUT

ETHNICS, POLITICS AND PSYCHOMETRICS IN DETECTING IRREGULARITIES IN SCHOOL DISTRICTS' TEST PROGRAMS (AERA/NCME SYMPOSIUM)

In complying with these demanding charges, the Testing Section of the Cincinnati Public Schools developed an administrative procedure detailing how to confirm exceptionally high test results. One highlight of this procedure entails the re-testing of schools that exceed by plus or minus 15% of students at and above national norm for the two previous years. Additionally, this procedure calls for the review of these results collectively by three people, an Area Director, Principal of the school where the results deviated and the Director of the Testing Section. The purpose of this review is to try to determine, ex post facto , plausible explanations for the deviations. If the panel of three all agree to the explanation given by the school principal, re-testing will not occur. On the other hand, if no plausible or sound rationale is forthcoming, re-testing will be done.

It should be noted that this procedure was followed to the letter for two years (1983 and 1984), and out of approximately 15 schools identified, only one in

each of these years was re-tested. In the 1984-85 school year, our Assistant Superintendent of Curriculum and Instruction decided to dispense with the review panel and re-test any or all schools where the 15% deviation occurred. The reasons for the change in procedure were twofold: (a) In the 1984-85 school year we would no longer be piloting a Design for School Excellence, but we would be fully implementing the Design, (b) We would be removing any bias in selection of schools.

In April of 1985, six schools out of a possible 80 schools were identified for re-testing. Principals were individually contacted by the Director of Testing to inform them that they would be re-tested. Schedules were prepared and a cadre of examiners/observers was assembled to assist these schools in the administration of the re-test. At this point, I believe, an important caveat should be highlighted: schools were initially tested the last week of March, 1985; tests were scored and analyzed from the middle of April to the middle of May, 1985. Principals were informed of the re-testing the third week in May. Re-testing actually occurred the last week in May, with Memorial Day being the first Monday of that week. I'm sure you would agree, this was not the best time to test; consequently, this turned out to be a very sore point with the schools involved.

With all parties informed, people assembled and tests packaged, the re-testing took place. Following are the major findings and results of this re-testing.

Results/Findings

The overall results can best be summarized in these words--"We survived the 1985 CAT re-testing."

Three statistical operations were performed on the test data for each of the six schools in question. They were:

1. Chi Square (x^2) was used by looking at the percent of students at and above norm in 1985 first test versus the 1985 results from the second test for each of the individual schools as well as using percent of students below norm.

2. A "t" test using Normal Curve Equivalents (NCEs) based on the school's average NCEs from the two tests was used to determine significance.

3. A "t" test was also performed by NCEs by grade in each of the schools

Specific findings from each of these methodologies and what was done with the results follows:

1. Chi Square (x^2)

Percent of students at and above norm was used as the observed frequencies for both testing situations and percent below norm was also used in the observation cells. The expected cell frequencies were computed from the row and column totals (marginals). Individual student data by school was used in the analysis. Results of utilizing this technique indicated that, at the school unit level, the re-test results significantly differed ($<.01$) in all six schools. Furthermore, overall results for each of the schools were lower than the original test results. Because it was felt this analysis was too general, i.e., no individual grades where much deviation was present and not taken into consideration, these results were not used to determine which results would be used i.e., first test or second test.

2. "t" Test -- Total School Normal Curve Equivalents (NCEs)

The next analysis that was performed involved looking at the average school unit NCEs from the first test to second test to determine if there was a significant difference in results obtained from each testing. Results showed, of the six schools re-tested, five out of the six displayed significant differences with the second test results, and for all six cases, the NCE results were more negative than the first. Like the x^2 test mentioned above, results from this analysis were also not used because of the little attention given to by-grade variations.

3. "t" Test -- By "Grade" Normal Curve Equivalents (NCEs)

In an effort to look at grade variations and to try and ferret out where significant change within a school actually took place, a "t" on average NCEs by grade from first testing to second testing was used. For the following reasons it was the results from this statistical analysis that were followed in determining which results (i.e., first test or second test) would "stand up" for the current year's (1985) test results:

a. Why criticize or praise the entire school if criticism or praise could be attributable to certain grade(s)?

b. The N's would be large enough to allow for statistical analysis.

c. By looking at a grade level aggregation of data, no specific teacher would be identified for either reward or punishment (as opposed to going to even a lower level of aggregation of classroom/teacher analysis).

Because results of this analysis were used together with percent of students at and above norm, a more detailed description of these findings will now be presented.

Based on this analysis, the obtained scores from the second testing either improved or were statistically the same as the first testing in well over half (60%) of the re-test situations. In the remaining situations substantial achievement differences were observed. Before going any further let me define the term "situation." In all there were 46 grades across schools and subject matters (reading and math) being tested and analyzed. Schematically this is represented below as Figure 1.

Figure 1: Testing Situations

School	Test	Grade 1	2	3	4	5	6
A - Primary Grades 1 - 3	Reading	X	X	X	(No grades 4, 5, 6)		
	Math	X	X	X			
B - Elementary Grades 1 - 6	Reading	X	X	X	X	X	X
	Math	X	X	X	X	X	X
C - Elementary Grades 1 - 6	Reading	(Did not exceed + or - 15%; did not test)					
	Math	X	X	X	X	X	X
D - Elementary Grades 1 - 5	Reading	X	X	X	X	X	(No grade 6)
	Math	X	X	X	X	X	
E - Elementary Grades 1 - 6	Reading	X	X	X	X	X	X
	Math	(Did not exceed + or - 15%; did not test)					
F - Elementary Grades 1 - 6	Reading	(Did not exceed + or - 15%; did not test)					
	Math	X	X	X	X	X	X

Note: X = Area re-tested (test situations = 46)

For this analysis, like the Chi Square analyses described previously, the .01 confidence level was established to determine significant change. If significant NCE differences existed, and the percent at and above norm differences were beyond plus or minus 5%, then results of the second testing were used. The plus or minus 5% was an error factor built in to allow for the effects of re-testing. In other words, if the differences between the percent of students at and above norm by grade from the first test to the second test were within a range of plus or minus 5%, the results of the first test were accepted regardless of the level of significance obtained.

Gains and Losses

The results by way of percent of students at and above norm are given in Table 1 for both initial test results (first time) and second test results.

Salient findings reveal as stated previously, in 27 out of the 46 (60%) testing situations, the re-test confirmed or upheld the results of the first testing. Therefore, results of the first testing were used and reported for that particular school/grade result in 1985. Based on absolute differences, 10 out of 46 situations (20%) actually improved. Schools identified as C and B in the math area in Table 1, had the most consistent test results from first test to second test. In reading, only school E had anything close to consistency, while school A obtained the most inconsistent results. For school A the results of the second test were used in all situations.

Table 1. Test and Re-test Results, 1985: Percent of Students At and Above Norm in Reading and Mathematics.

School	Test	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Total
A	*R_1	85.9%	76.8%	89.8%				82.5%
	R_2	73.1	54.4	72.7				66.7
	***D	-12.8②	-22.4②	-17.1②				-15.8
	**M_1	85.9	76.8	89.8				83.7
	M_2	68.1	49.9	72.7				62.8
	D	-17.8②	-26.9②	-17.1②				-20.9
B	R_1	74.1	72.1	28.3	44.2	48.1	44.4	52.6
	R_2	72.2	54.3	26.2	30.8	24.5	32.6	40.8
	D	- 1.9①	-17.8②	- 2.1①	-13.4②	-23.6②	-11.8②	-11.8
	M_1	63.2	53.7	36.2	31.7	39.2	34.1	43.8
	M_2	44.4	31.4	41.9	29.7	44.9	36.6	39.0
	D	-18.8②	-22.3②	+ 5.7①	- 2.0①	+ 5.7①	+ 2.5①	- 4.8
C	M_1	88.4	91.9	87.2	92.3	77.1	82.1	86.6
	M_2	80.0	97.1	94.3	80.6	71.1	82.9	84.0
	D	- 8.4①	+ 5.2①	+ 7.1①	-11.7②	- 6.0①	+ 0.8①	- 2.6
D	R_1	57.4	54.5	37.5	66.1	76.9		55.8
	R_2	54.7	51.0	31.0	36.0	83.3		46.5
	D	- 2.7①	- 3.5①	- 6.5②	-30.1②	+ 6.4①		- 9.3
	M_1	55.9	63.6	45.8	53.6	84.6		56.7
	M_2	45.2	45.1	44.2	48.9	83.3		47.9
	D	-10.7①	-18.5②	- 1.6①	- 4.7①	- 1.3①		- 8.8
E	R_1	63.5	85.7	48.5	50.8	35.4	28.6	56.3
	R_2	53.6	87.3	49.1	38.1	31.9	28.6	52.0
	D	- 9.9②	+ 1.6①	+ 0.6①	-12.7②	- 3.5①	NC ①	- 4.3
F	M_1	38.6	63.3	43.5	45.1	73.7	60.0	51.1
	M_2	21.2	57.8	44.4	42.0	63.2	54.1	44.4
	D	-17.4②	- 5.5①	+ 0.9①	- 3.1①	-10.5①	- 5.9①	- 6.7

[1] Results used = first test
[2] Results used = second test--significant at the <.05 level

* R_1 = Reading First Test R_2 = Reading Second Test
** M_1 = Mathematics First Test M_2 = Mathematics Second Test
*** D = Difference between R_1 and R_2 and/or M_1 and M_2

Probably many of you are now thinking either "so what" or "what did you do with these findings?" Let me elaborate. Each principal was brought into my office and all the information presented to you today was gone over very, very carefully with them. Additionally, I made it very clear to the principals in those schools where the re-test did not support the results of the first test that the results of the first test were invalid and re-test scores had been substituted. I also informed them that they should take careful measures to monitor these situations next year. I made it abundantly clear to them that though we remain cautious about the substantial differences observed, and though we may be implying cheating, what we really were endeavoring to do was to validate and point up differences if they existed, and at the same time, not suggesting what may have caused these differences or that cheating took place. In fact, if I were to suggest what occurred, my feeling would be one or the other or a combination of the following:

1. Unconscious misadministration of the test, i.e., allowing for too much time for students to take the test or answering questions that would direct children to an answer.

2. Being too meticulous in scheduling or going to the extreme of protecting children from fatigue. In other words, only scheduling one subtest per day, even though that particular subtest only consisted of 11 items, and total time for administration had to be no longer than 10 or 12 minutes. Let me digress for a moment on this issue of test scheduling. Test publishers and their respective management guides and administration manuals are, in my opinion, not emphatic enough in spelling out how much of a test should be completed in a given day. Admittedly, they do a very good job on giving time limits and number of days, but they do not specify what is an adequate portion of the test to be covered in a given day.

3. Direct instruction toward the test--not teaching the test itelf--with a heavy emphasis on teaching skills and objectives of the test, in some schools, one and two months prior to when the test is given.

4. Prompted to teach and the teaching of "Test Wiseness Skills" with use of supplementary materials such as: Scoring High, Scoring High on the California Achievement Test, film strips, and much inservice in the area by my staff members who are Local School Evaluators. (A problem with respect to these latter two statements--Are we crossing the fine line into irregularities in testing or are we within bounds? I can't honestly say!)

5. Erasures When reviewing the first set of data of these six schools, we initially noticed in one school what appeared to us to be an inordinate amount of erasures. Consequently, at the outset, we thought this to be the problem. Since this analysis is not only interesting, but leaves questions still in our minds, I would like to elaborate a little more on this topic.

The informal design called for looking at not only the number of erasures/answers changed, but the percent of changed answers to a correct answer. Additionally, we also looked at three control schools in order to make an informed judgment. Findings from this mini-research study are given in Table 2.

Table 2. Control Group Versus Experimental Group--Erasures/Answer Changes, Spring, 1985.

Erasures/Answer Changes	Experimental Group	Control Group
Number of possible changes	112,542	64,761
Number of changes to positive	4,706	2,013
Percent positive changes	62%	65%
Number of changes overall	7,538	3,106
Percent changes overall	6.7%	4.8%

As Table 2 indicates, there was a higher proportion of erasures made overall by the experimental group (schools involved in re-testing) versus the control group of schools (randomly selected schools not participating in re-testing). The percentage of change/erasures made overall by the re-tested schools was 6.7% in contrast to 4.8% for the non re-tested schools.

Another finding, and perhaps the more critical point to be made, relates to the fact that a higher percentage of non re-tested student erasures (65%) resulted in positive changes to the "correct" answer compared to the re-tested students (62%).

Because the results of this analysis are not conclusive, no statement at

this time can be made as to cheating due to answer changing. Additionally, what also taints the analysis involves a checklist that I provide teachers for administering tests. This checklist includes the following directions:

... check conditions of answer forms

_____marks are dark

_____ERASURES ARE CLEAN

_____check name grids and student I.D. numbers

You can be sure that we will be checking this area (erasures) this year and you can be doubly sure that I will no longer be offering this same set of directions.

Discussion

Reactions to our efforts have ranged from extremely negative to extremely positive. Prior to the actual re-testing of schools, all reactions about the concept from the field and community were negative. However, after having gone through the re-testing, I would say twenty-five percent of the comments we received were positive. For example, our two major newspapers came out positively toward our efforts. A third newspaper, however, was quite negative in charging us with racial discrimination against poor, black schools which, by the way, four out of the six were.

Some random comments and accusations made by various groups and/or individuals and publications follow:

"...of the six schools being re-tested, five have predominately Black populations." (There were only four)

"...our initial concern is that such re-testing, intentionally or not, is automatically received by the community as to cheating on the original test."

"...you're only re-testing schools that have female principals." (one school had a male)

"...you're only re-testing in schools that have an Open Court reading program."

"...the teachers at my school are outraged that students were given the CAT the second time with such short notice and a poor testing environment."

"...we think it is very unfair that our children were expected to do their best the week before school is out."

"...our main concern is how you can consider the second testing situation as proof of validity under such weak circumstances."

"...as parents and LSAC (Local School Advisory Council) members of our school, we are writing to protest the imposition of a second round of standardized tests at the school this year. In our opinion, the first round of tests is more valid while the re-test was characterized by adverse conditions and caused an unnecessary disruption of the curriculum."

"...the business of a school is teaching: testing is merely an evaluative method."

"...Mr. Assistant Superintendent, it is my understanding that you are the one that made the decision to re-test our children with the CAT."

"...as you can see, validation DID NOT occur. While I do recognize the desire by Testing/Evaluation Department for the most accurate scores, and the friendliness and helpfulness of the observers and Dr. Barbadora, I strongly recommend that:
Instead of destroying morale of the teachers, parents and children, I appeal to you to schedule at least one monitor to roam our building during 1986 testing or test earlier in the year to provide appropriate time and atmosphere for re-testing."

"...school proud of success on second retest"

"...Six Cincinnati Public Schools have survived 'the great retest of '85' and survived, by all indications, with flying colors."

"...the system takes its testing seriously and works to make the scores accurately reflect the growth of its pupils. To make sure the results are statistically valid, administrators re-test students at schools which show an unusual change in their scores over previous years."

"...the testing program should not be seen as a competition, but as a tool for evaluating performance of a multi-million dollar annual investment in our future."

Regarding the previous quotations, we are now in the process of responding to each and every one. (By the way, if you would like to assist me in this endeavor, please see me after the meeting.) On a more serious note, these are

highlighted to give you somewhat of a sampling of what you are also in for and the politics that can be involved.

In closing, I personally feel re-testing has to and must continue to be done--if necessary. To quote again the words of our late superintendent, "we must make it very clear to principals, teachers and community, that we will leave no stone unturned to protect the integrity of the testing process." Further, if blame has to be cast, I do not want to stand here before you to have you think that I, personally, or the adminstration is without fault. We know better. We are part of the process and problem and, hopefully, the solution. To begin ameliorating the problems from our end, one of the changes to be made for next year's testing is a combination of inservice and better procedures regarding test administration. It can be assured that within the constraints of resources (both human and financial) and time, the Cincinnati Public Schools will continually strive to assure not only accurate test data, but accurate data on all output indicators of the Design for School Excellence and all other data as well. James Najeeb Jacobs strived for nothing less and we will continue to keep and implement this most gratifying high ideal.